Copyright © 2000 by the Genetics Society of America

Bayesian Mapping of Quantitative Trait Loci for Complex Binary Traits

Nengjun Yi and Shizhong Xu

Department of Botany and Plant Sciences, University of California, Riverside, California 92521-0124

Manuscript received August 23, 1999
Accepted for publication March 6, 2000

ABSTRACT

A complex binary trait is a character that has a dichotomous expression but with a polygenic genetic background. Mapping quantitative trait loci (QTL) for such traits is difficult because of the discrete nature and the reduced variation in the phenotypic distribution. Bayesian statistics are proved to be a powerful tool for solving complicated genetic problems, such as multiple QTL with nonadditive effects, and have been successfully applied to QTL mapping for continuous traits. In this study, we show that Bayesian statistics are particularly useful for mapping QTL for complex binary traits. We model the binary trait under the classical threshold model of quantitative genetics. The Bayesian mapping statistics are developed on the basis of the idea of data augmentation. This treatment allows an easy way to generate the value of a hypothetical underlying variable (called the liability) and a threshold, which in turn allow the use of existing Bayesian statistics. The reversible jump Markov chain Monte Carlo algorithm is used to simulate the posterior samples of all unknowns, including the number of QTL, the locations and effects of identified QTL, genotypes of each individual at both the QTL and markers, and eventually the liability of each individual. The Bayesian mapping ends with an estimation of the joint posterior distribution of the number of QTL and the locations and effects of the identified QTL. Utilities of the method are demonstrated using a simulated outbred full-sib family. A computer program written in FORTRAN language is freely available on request.

THE overwhelming amount of molecular data provides a large opportunity to locate genes controlling the expression of quantitative traits. Currently, a variety of statistical methods are available for mapping quantitative trait loci (QTL). Early methods of QTL mapping were developed on the basis of the maximum-likelihood or least-squares method under a single QTL model (*e.g.*, Lander and Botstein 1989; Haley and Knott 1992). Yet it is now known that when multiple QTL are present in the same linkage group, the single QTL model can lead to biased estimates of QTL positions and effects (*e.g.*, Haley and Knott 1992). In theory, effects of multiple QTL can be simultaneously included in the model, but this is difficult to implement in practice because even the number of QTL is an unknown parameter. Jansen (1993) and Zeng (1994) developed the idea of composite interval mapping in which mapping in a particular interval is combined with multiple regression on markers in other chromosomal regions to absorb effects of other QTL. Recently, Kao *et al.* (1999) developed a multiple interval mapping (MIM) approach particularly designed for mapping multiple QTL. All these approaches provide only point estimates for number, locations, and effects of QTL. The critical values for significance tests and interval estimates of the parameters have to be established using a repeated sampling technique, *e.g.*, a permutation test (Churchill and Doerge 1994) or bootstrapping (Visscher *et al.* 1996b).

Bayesian methods of QTL mapping have been developed, in particular, for detection of multiple QTL (Satagopan and Yandell 1996; Satagopan *et al.* 1996; Heath 1997; Uimari and Hoeschele 1997; Stephens and Fisch 1998; Sillanpää and Arjas 1998, 1999). In the Bayesian analysis, rather than maximizing a likelihood function, inferences are based on the joint posterior distribution of all unknown variables given the prior distribution of all unknowns and the observed data. The introduction of iterative simulation methods, such as the data augmentation and the more general Markov chain Monte Carlo (MCMC) algorithm (Tanner and Wong 1987; Gelfand and Smith 1990), which provide a Monte Carlo approximation to the required multiple integration, has brought the Bayesian method into the mainstream of QTL mapping. For complicated pedigree data, as commonly seen in animal breeding, Bayesian QTL mapping was demonstrated by Hoeschele and VanRanden (1993a,b), implemented via MCMC by Thaller and Hoeschele (1996) for single markers, by Uimari *et al.* (1996) for multiple linked markers, and by Uimari and Hoeschele (1997) for multiple linked QTL. In plants, Bayesian mapping has been seen in Satagopan *et al.* (1996), where a prespecified number of QTL were assumed first, and then a Bayes factor approach was used to decide the most probable number of QTL. Using the reversible jump MCMC (Green 1995), researchers can even treat the number of QTL

Corresponding author: Shizhong Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521.
E-mail: xu@genetics.ucr.edu

as an unknown variable and generate its posterior distribution (Satagopan and Yandell 1996; Heath 1997; Stephens and Fisch 1998; Sillanpää and Arjas 1998, 1999). This full Bayesian treatment has further revolutionized QTL mapping and opened a new horizon in quantitative genetics.

Almost all the Bayesian mapping methods mentioned above are designed for normally distributed traits. Many traits of biological interest and/or economical importance, however, show a dichotomous or binary phenotype, but are not inherited in a simple Mendelian manner. The genetic architectures of these characters are generally complex, involving multiple interacting genetic factors. Furthermore, the expression of the phenotype is often sensitive to environmental factors. As a consequence, these traits are usually explained by the concept of the threshold model (Falconer and Mackay 1996; Lynch and Walsh 1998), which assumes a latent continuous variable (called the liability) underlying a binary trait. The binary phenotype and the continuous liability are linked through a fixed threshold. One can treat the liability as an unobservable quantitative trait. Genes controlling complex binary traits can be treated as quantitative trait loci and handled using the QTL mapping approach.

QTL mapping for the liability of a binary trait is more complicated than for a regular quantitative trait. Although considerable progress has been made over the past few years, the development of new statistical methodology for binary traits still poses a great challenge. In human linkage studies, a nonparametric approach has been proposed (Kruglyak and Lander 1995). Under the threshold model, parametric methods of QTL mapping based on a generalized linear model (GLM) have been developed in line crosses (Hackett and Weller 1995; Xu and Atchley 1996; Visscher *et al.* 1996a; Rebai 1997; Rao and Xu 1998; Xu *et al.* 1998). Yi and Xu (1999a,b) recently developed a random model approach to QTL mapping for complex binary traits. Because of the additional complexity added between the phenotype and the liability, these methods were developed on the basis of either a single QTL model or with some approximation. A Bayesian mapping method has not been available for binary traits and such a method is ideal for handling problems with this level of complexity. Therefore, the purpose of this study is to explore the application of Bayesian mapping to binary and multiple-ordered categorical traits.

STATISTICAL METHODS

The threshold model and liability: Let s_i and y_i ($i = 1, \cdots, n$) be the binary phenotype and the underlying liability, respectively, of the ith individual in a full-sib family of interest. We are interested in developing a QTL mapping algorithm in a full-sib family because a full-sib family represents the simplest form of an outbred or open-pollinated population, and the method can be easily extended to natural populations. The threshold model assumes that there is a fixed threshold in the scale of liability, t, which determines the binary phenotype of an individual by comparing y_i with t. If $y_i > t$, we assign $s_i = 1$, and otherwise $s_i = 0$. The liability y_i is treated as a usual quantitative character and is thus described by the linear model,

$$y_i = \mathbf{X}_i^{\mathrm{T}}\beta + \sum_{j=1}^{l} \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{H}\gamma_j + \varepsilon_i, \qquad (1)$$

where β is a $p \times 1$ vector of covariate effects (including the overall mean), which relate y_i via a known incidence vector $\mathbf{X}_i$; ε_i is the residual effect (including the environmental error) distributed as $N(0, \sigma_\varepsilon^2)$; l is the number of QTL affecting the liability on all chromosomes; $\gamma_j = (\alpha_j^{\mathrm{m}}, \alpha_j^{\mathrm{f}}, \delta_j)^{\mathrm{T}}$ is a vector of genetic effects of the jth QTL with α_j^{m} and α_j^{f} being the maternally and paternally inherited allelic effects, respectively, and δ_j the dominance effect; $\mathbf{Z}_{ij} = (z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4})^{\mathrm{T}}$ are indicators for the four possible ordered genotypes and defined as $z_{ijk} = 1$ if the kth genotype is observed and $z_{ijk} = 0$ otherwise; and $\mathbf{H} = (\mathbf{H}_{\mathrm{m}}\ \mathbf{H}_{\mathrm{f}}\ \mathbf{H}_{\delta})$, where $\mathbf{H}_{\mathrm{m}} = (1\ 1\ -1\ -1)^{\mathrm{T}}$, $\mathbf{H}_{\mathrm{f}} = (1\ -1\ 1\ -1)^{\mathrm{T}}$, and $\mathbf{H}_{\delta} = (1\ -1\ -1\ 1)^{\mathrm{T}}$ represent the linear contrasts converting the three genetic effects into the genotypic values of the four genotypes. The threshold model is overparameterized so that some constraints must be superimposed. As usual, we take $\sigma_\varepsilon^2 = 1$ and $t = 0$ (Albert and Chib 1993; Sorenson *et al.* 1995). Note that the four genotypes in the progeny are ordered as $\{A_1A_3, A_1A_4, A_2A_3, A_2A_4\}$ given the genotypes of the parents being A_1A_2 and A_3A_4.

The observables are the binary phenotypic values $\mathbf{S} = \{s_i\}_{i=1}^n$, the covariates, and the marker data. The locations of markers on chromosomes are known *a priori*. Marker linkage phases in the parents are assumed known once they are inferred from marker data of the progeny. When the family size is small, inference of marker linkage phases can be subject to error and the linkage phases should be sampled along with other unknowns (Sillanpää and Arjas 1999). The observed marker genotypes in some individuals may not be fully informative and the patterns of allelic inheritance of such markers are also unknown. The list of unobservables includes the liability $\mathbf{Y} = \{y_i\}_{i=1}^n$, the number of QTL and their locations $\lambda = \{\lambda_j\}_{j=1}^l$, the complete marker genotypes $\mathbf{M} = \{\mathbf{M}_{ik}\}$, the QTL genotypes $\mathbf{Z} = \{\mathbf{Z}_{ij}\}$, and the model effects $\theta = (\beta^{\mathrm{T}}, \gamma_1^{\mathrm{T}}, \cdots, \gamma_l^{\mathrm{T}})^{\mathrm{T}}$, where λ_j denotes the distance of the jth QTL from one end of the corresponding chromosome, $\mathbf{M}_{ik}$ and $\mathbf{Z}_{ij}$ denote the kth marker genotype and the jth QTL genotype, respectively, for the ith individual.

From Bayes' theorem, the joint posterior density of the unobservables, $\{\mathbf{Y}, l, \lambda, \theta, \mathbf{M}, \mathbf{Z}\}$, given the binary data $\mathbf{S}$, is

$$p(\mathbf{Y}, l, \lambda, \mathbf{M}, \mathbf{Z}, \theta|\mathbf{S}) \propto p(\mathbf{S}|\mathbf{Y}, l, \lambda, \mathbf{Z}, \theta)p(\mathbf{Y}|l, \lambda, \mathbf{Z}, \theta) \times p(\mathbf{Z}|l, \lambda, \mathbf{M})p(\mathbf{M})p(l, \lambda, \theta). \quad (2)$$

Here, we have suppressed the notation for conditional on the observed marker data. The first term in (2) is the conditional distribution of the data given all the unknowns. It is given by Albert and Chib (1993) and Sorensen *et al.* (1995) and has the form of

$$p(\mathbf{S}|\mathbf{Y}, l, \lambda, \mathbf{Z}, \theta) = \prod_{i=1}^{n} p(s_i|y_i) = \prod_{i=1}^{n} \{1(y_i > 0)1(s_i = 1) + 1(y_i < 0)1(s_i = 0)\}, \quad (3)$$

where $1(X \in A)$ is the indicator function, taking the value of 1 if X is contained in A, and 0 otherwise. Note that $p(\mathbf{S}|\mathbf{Y}, l, \lambda, \mathbf{Z}, \theta) = p(\mathbf{S}|\mathbf{Y})$ because $\mathbf{S}$ is solely determined by $\mathbf{Y}$. The second term in (2) is the conditional distribution of the liability given all other unknowns. Because liabilities of individuals are normally distributed and independent of each other given other unknowns, we have

$$p(\mathbf{Y}|l, \lambda, \mathbf{Z}, \theta) = \prod_{i=1}^{n} p(y_i|l, \lambda, \mathbf{Z}, \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y_i - \mathbf{X}_i^{\mathrm{T}}\beta - \sum_{j=1}^{l} \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{H}\gamma_i)^2\right\}. \quad (4)$$

The next term in (2) is $p(\mathbf{Z}|l, \lambda, \mathbf{M})$, which is the QTL genotype distribution conditional on the number and locations of QTL and the complete marker genotypes. $p(\mathbf{M})$ is the complete marker genotype distribution conditional on observed marker information (recall that markers can be partially informative). $p(\mathbf{Z}|l, \lambda, \mathbf{M})$ and $p(\mathbf{M})$ are derived later. Finally, the last term in (2), $p(l, \lambda, \theta)$, is the joint prior distribution of l, λ, and θ, the parameters of interest.

Prior distributions: Assuming prior independence of the locations and effects of QTL, the joint prior density $p(l, \lambda, \theta)$ can be factored into the following product:

$$p(l, \lambda, \theta) = p(l)p(\lambda|l)p(\theta|l) = p(l)p(\beta)\prod_{j=1}^{l}[p(\lambda_j)p(\alpha_j^{\mathrm{m}})p(\alpha_j^{\mathrm{f}})p(\delta_j)]. \quad (5)$$

As in Sillanpää and Arjas (1998, 1999), the prior distribution of l (the number of QTL) is assumed to be truncated Poisson with mean μ and the maximum number L. When no information regarding the locations is available, the prior probability that a QTL is on a chromosome is proportional to the length of the chromosome. Within a chromosome, each QTL has a uniform distribution of residing at any location on that chromosome. We use diffuse normal priors for all regression parameters, including the QTL effects and the fixed effects, *e.g.*, the effect of sex and experimental sites.

In updating marker genotypes, we use the prior probabilities of the complete marker genotype at each marker locus for each individual. These prior probabilities are calculated using the simplified multipoint method of Rao and Xu (1998) in which genotypes of all markers, including the marker in question, are used to infer the prior probabilities of the allelic inherence of the current marker. This treatment guarantees that a marker genotype sampled is compatible with observed data. When a marker is already fully informative, each genotype is uniquely identified, and the multipoint prior will automatically force the marker locus not to be updated. When the marker information content is low, using the multipoint prior can increase the efficiency of MCMC compared with the two-point prior (using flanking markers only).

Conditional posterior distributions: To implement the MCMC algorithm, conditional posterior distributions of the unknowns are needed. From the joint posterior density given in (2), the conditional posterior density of each unknown can be derived by treating other elements in the list of unknowns as constants and selecting the terms involving the item of interest. When this leads to the kernel of a standard density, *e.g.*, normal distribution, Gibbs sampling is applied to draw samples for that distribution. Otherwise, sampling needs to be done by using other techniques, *e.g.*, the Metropolis-Hastings algorithm.

To facilitate the development of the Gibbs sampler, the unobserved liability is included as an unknown nuisance parameter in the model. This approach, known as data augmentation, has been used in Bayesian analysis under the polygenic model of threshold traits (Sorensen *et al.* 1995), but not in the context of QTL mapping. The conditional posterior distribution of the underlying variable y_i given the binary phenotype is a truncated normal. The probability density of this truncated normal distribution is given in appendix a. The algorithm for simulating a truncated normal variable is given by Devroye (1986).

Given the liability $\mathbf{Y}$, the number l, and genotype $\mathbf{Z}$ of QTL, the posterior distributions for the elements of θ can be derived using the standard linear model theory under a normal distribution. If normal priors are chosen for the regression coefficients θ, these posterior distributions are also normal with a mean and variance as given in appendix a. Therefore, the augmentation approach allows the use of existing MCMC algorithms for normally distributed variables. Conditional on all other parameters, marker and QTL genotypes of each individual are independent and therefore can be updated locus by locus and individual by individual (Jansen *et al.* 1998). Expressions of the conditional posterior probabilities are derived using Bayes' theorem, which is given in appendix a. Unfortunately, there are no explicit expressions for the conditional posterior distributions of the number l and locations λ of QTL. Updating of l and λ

must be achieved using the Metropolis-Hastings algorithm.

MCMC algorithm: A Markov chain Monte Carlo method is used to generate the joint posterior distribution of all unknowns given in Equation 2. The idea of MCMC is to simulate a random walk in the space of all unknowns that converges to a stationary distribution (Gelman *et al.* 1995). The stationary distribution represents the posterior distribution of the unknowns. Various approaches have been suggested to conduct MCMC. Two commonly used approaches are the Gibbs sampler and the Metropolis-Hastings algorithms. The reversible jump MCMC is an extension of the Metropolis-Hastings sampler, permitting posterior samples to be collected from posterior distributions with varying dimensions (Green 1995; Richardson and Green 1997). The proposed MCMC algorithm is carried out in an alternating conditional sampling fashion. In other words, each iteration of the sampling cycles through all elements of unknowns $\{\mathbf{Y}, \theta, l, \lambda, \mathbf{M}, \mathbf{Z}\}$ represents the drawing of each unknown conditional on the current values of all other unknowns. The algorithm starts from an initial point $(\mathbf{Y}^0, \theta^0, l^0, \lambda^0, \mathbf{M}^0, \mathbf{Z}^0)$ and proceeds to modify each of the unknowns in turn. To set initial values for $\mathbf{M}^0$ and $\mathbf{Z}^0$, we first calculate the probabilities of marker and QTL genotypes using the simplified multipoint method (Rao and Xu 1998) and then sample randomly a value from the probability distribution as the initial points of marker and QTL genotypes. Given the initial values of $(\theta^0, l^0, \lambda^0, \mathbf{M}^0, \mathbf{Z}^0)$, we generate $\mathbf{Y}^0$ from the corresponding truncated normal distributions. The Gibbs sampler approach is adopted to update $\mathbf{Y}$, θ, $\mathbf{M}$, and $\mathbf{Z}$ because the conditional posterior distributions of $\mathbf{Y}$, θ, $\mathbf{M}$, and $\mathbf{Z}$ have standard forms (appendix a). Updating λ is implemented using the Metropolis-Hastings algorithm. Updating of the QTL number l requires a change in the dimension of the model and thus needs a reversible jump step. This has been accomplished by Sillanpää and Arjas (1998, 1999) and is directly adopted in this study. The MCMC process is briefly summarized as follows:

Step 1: Update the liability $\mathbf{Y}$ individual by individual using (A1) and (A2);
Step 2: Update coefficients β and $\{\gamma_i\}_{i=1}^{l}$ using (A3)–(A5);
Step 3: Update the number l of QTL and their locations λ.
Step 4: Update the QTL genotypes individual by individual and locus by locus using (A6);
Step 5: Update marker genotypes individual by individual and locus by locus using (A7).

In step 3, we make a random choice among three move types: (1) modify the QTL locations; (2) add one new QTL to the model; and (3) delete one QTL from the model, with probabilities p_m, p_a, and $p_d = 1 - p_m - p_a$, respectively. Of course, $p_a = 0$ if $l = L$ and $p_d = 0$ if $l = 0$, and otherwise we choose $p_m = p_a = p_d = \frac{1}{3}$, for $0 < l < L$.

As in Sillanpää and Arjas (1998, 1999), we do not fix the order of QTL when updating the QTL locations. Elements of λ are modified one at a time using the Metropolis-Hastings algorithm. For the jth QTL, a proposal λ_j^{new} is sampled from a uniform distribution in the neighborhood of the previous value λ_j. The probability that the proposal is accepted is given in appendix b. If the proposal is not accepted, the state remains unchanged, and the algorithm proceeds to update the next QTL location.

To add a QTL, we must sample a new location, new genetic effects, and new QTL genotypes for each individual. First, we sample a chromosome with a probability proportional to the length of the chromosome. Once a chromosome is chosen, a location of the new QTL λ_{l+1} is proposed from the uniform density on the sampled chromosome. We then sample QTL genotypes for the newly added QTL for each individual from $p(\mathbf{Z}_i^{new}|\lambda_{l+1}, \mathbf{Z}_i^l, \mathbf{Z}_i^r)$, where $\mathbf{Z}_i^l$ and $\mathbf{Z}_i^r$ are the left and right flanking genotype of the ith individual for the proposed QTL. The flanking genotypes can be the genotypes of markers or QTL, whichever are closer to the new QTL position. Finally, new QTL effects are drawn from a normal density $N(0, \sigma^2)$, where σ^2 is a prespecified constant. We then calculate the acceptance probability using equations given in appendix b and carry out a Metropolis-Hastings step. If the proposal is accepted, then we add a QTL to the model at the new location and the genotypes and the effects are all accepted simultaneously. Otherwise, the state remains unchanged.

To delete a QTL, an existing QTL is selected at random and the relevant acceptance probability is calculated. The form of the acceptance probability is also given in appendix b. If the proposal is accepted, we delete a QTL from the model. Otherwise, the QTL number remains unchanged.

SIMULATION STUDIES

Designs of simulations: The Bayesian method was evaluated empirically by analyzing simulated data. We simulated two chromosomes of length 100 cM and 70 cM, respectively. Eleven and 8 codominant markers were respectively placed on the two chromosomes with a marker distance of 10 cM. Four equally frequent alleles were simulated at each marker locus. Marker genotypes were observed for parents and all the offspring. In each design, we simulated an outbred full-sib family with 300 individuals. Three QTL were simulated to control the expression of a binary trait. There were exactly four alleles at each QTL (each parent has two distinguished alleles and the two parents are not related). Therefore, the QTL were fully informative. We used the size of the variance explained by each QTL to control the polymorphism of the QTL. If the variance is zero, the polymor-

TABLE 1

The true locations and genetic effects of the three simulated QTL

Design	Chromosome	Location (cM)	Paternal allelic effect (α_j^m)	Maternal allelic effect (α_j^f)	Dominance effect (δ_j)	Heritability
I	1	25	0.3162	0.3162	0.4472	0.20
	1	75	0.2236	0.2236	0.3162	0.10
	2	25	0.3162	0.3162	0.4472	0.20
II	1	25	0.1414	0.1414	0.2000	0.06
	1	75	0.1360	0.1360	0.1923	0.05
	2	25	0.1871	0.1871	0.2646	0.10

The effects are in the scale of liability for the binary data. The heritability is defined as proportion of the variance of liability explained by the locus of interest.

phism of the QTL does not mean anything, although the four alleles are distinguished. On the other hand, if the variance is very large, but all four alleles are identical, we get the equivalent result as zero variance. Therefore, we decided to control only the variance and let the QTL be fully informative. The locations and effects of the three simulated QTL are given in Table 1. The value of the liability of each individual took the sum of the overall mean, values of QTL additive and dominance effects, and an environmental error sampled from a standardized normal distribution. The observed binary phenotype was set to be $s_j = 1$ if the corresponding liability exceeds $t = 0$, and $s_j = 0$ otherwise. We designed two simulation experiments. In design I, the overall mean of the liability was set at 0.0, which generates a trait incidence of 50%. In design II, the mean was set at -0.95, leading to a trait incidence of 20%.

For comparison, the liability of each individual was also reported and analyzed as if it were observed. Therefore, we analyzed two data sets produced from the same group of sibs, the binary data and the normally distributed data. When the normally distributed data were analyzed, we used the same program by suppressing the liability-generating subroutine and replacing the liabilities by the reported normal observations. In addition, we added a subroutine to generate the environmental error because σ_ε^2 must be estimated instead of taking a value of unity. The residual variance at each cycle was updated from its conditional posterior distribution $p(\sigma_\varepsilon^2|\mathbf{Y}, l, \lambda, \mathbf{M}, \mathbf{Z}, \theta)$ in the MCMC algorithm. This conditional posterior distribution is an inverse gamma according to the standard linear model theory when the prior for σ_ε^2 is an inverse gamma distribution (*e.g.*, Gelman *et al.* 1995; Sorensen *et al.* 1995; Satagopan *et al.* 1996). For simplicity of programming, however, we simply used the Metropolis-Hastings (M-H) algorithm to generate σ_ε^2 under a flat prior.

We modified the maximum-likelihood (ML) method of Xu and Atchley (1996) so that it can handle the data with such a full-sib family structure and compared our Bayesian results with the ML analysis. The logistic function was replaced by the probit function. The ML was implemented using an EM algorithm derived by S. Xu (unpublished results).

For all MCMC analyses, the same initial values and priors were used. The initial value for the QTL number was set at 2 and the corresponding locations were at 50.0 cM of chromosome 1 and 40.0 cM of chromosome 2, respectively. The prior Poisson mean of the number of QTL was $\mu = 2$ and the maximum number of QTL was $L = 6$. The starting values for all regression parameters were 0.0. The priors for the regression parameters were normal with mean 0.0 and variance 10.0. The prior for the QTL locations was uniform over the whole genome. The tuning parameter of the proposal distribution for QTL locations was chosen to be 2.0 cM. Finally, the proposal distributions for the allelic and dominance effects were normal with means 0.0 and variance 1.0 in cases where the addition of a new QTL to the model was proposed.

In each of the MCMC analyses, we ran a single long chain with 10^6 cycles of simulations. The first 200 samples (burn-in period) were discarded and the chain was thinned (saved one iteration in every 50 cycles) to reduce serial correlation in the stored samples so that the total number of samples kept in the analysis was 2×10^4.

RESULTS

Before we present the result of Bayesian mapping, we first give the results of the ML analysis for the binary data. Figure 1 shows the likelihood profiles along the two chromosomes for both designs. For design I (Figure 1, a and b), we observed one peak at position 26 cM in chromosome 1, overlapping with the true location of the first simulated QTL (at 25 cM). The estimated effects of this QTL are $\hat{\alpha}_j^m = 0.3744$, $\hat{\alpha}_j^f = 0.4024$, and $\hat{\delta}_j = 0.3858$, very close to the true values. There is no evident peak in the neighborhood of the second QTL, although the test statistics are consistently high. This clearly demonstrates the limitation of the single QTL ML analysis. The second chromosome shows a clear peak at 22 cM for design I and the estimated effects of the identified

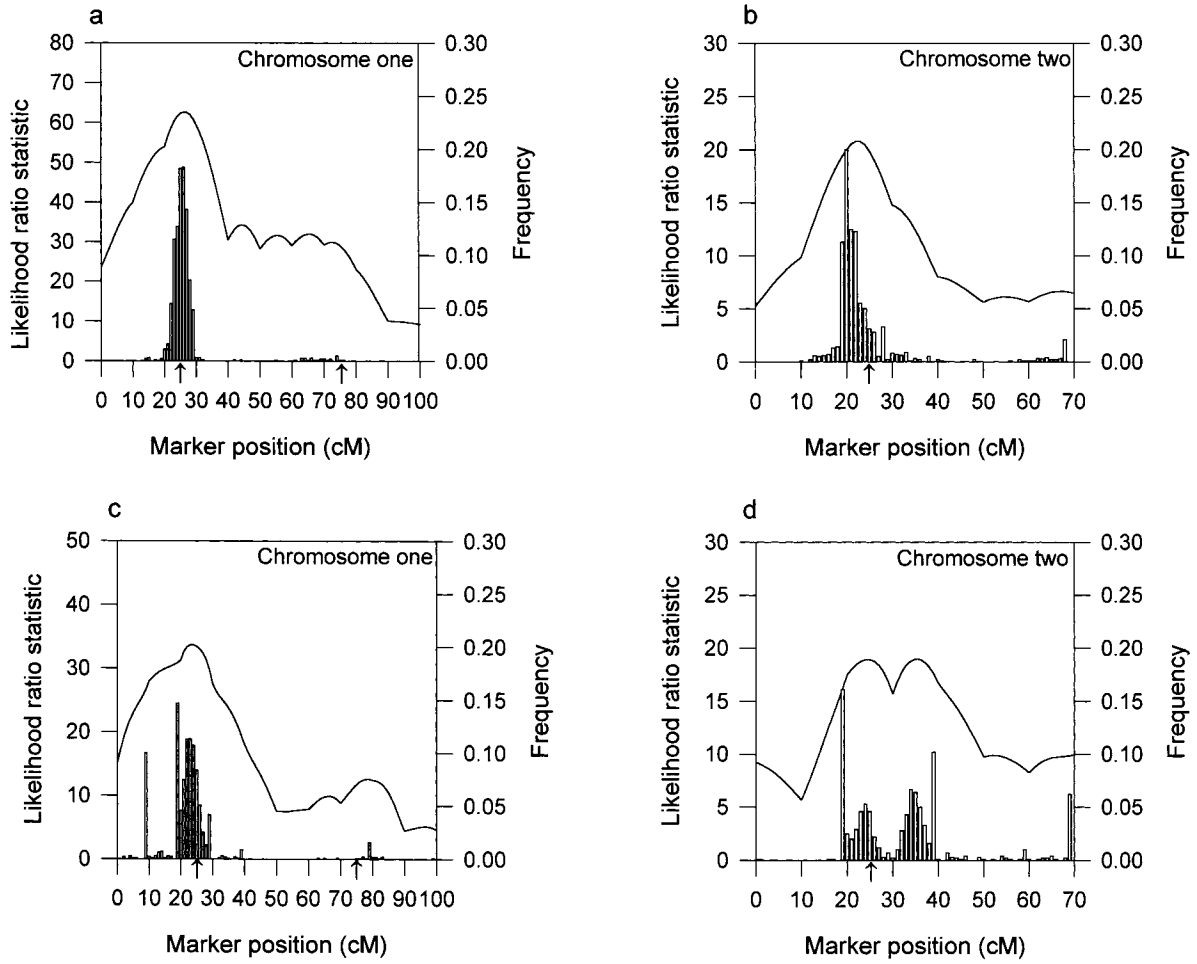

Figure 1.—Likelihood-ratio profiles of ML mapping and empirical distributions of the estimated QTL position obtained by 1000 bootstrap samples from the simulated binary data in design I (a and b) and design II (c and d). The solid curves are the likelihood-ratio profiles and the histograms are the bootstrap frequencies. The left *y*-axis corresponds to the likelihood-ratio statistic and the right *y*-axis corresponds to the bootstrap frequency. The true locations of the simulated QTL are indicated with an arrow (↑).

QTL are $\hat{\alpha}_j^m = 0.1008$, $\hat{\alpha}_j^f = 0.1654$, and $\hat{\delta}_j = 0.3893$. One should not expect the estimated values to be identical to the true values because this only represents the result of one random sample with 300 individuals.

For design II (Figure 1, c and d), a major peak was observed at 24 cM in chromosome 1 and the corresponding effects were estimated to be $\hat{\alpha}_j^m = 0.3193$, $\hat{\alpha}_j^f = 0.1873$, and $\hat{\delta}_j = 0.3230$. However, the second QTL in chromosome 1 remained undetected due to the low likelihood-ratio value (12.6101). For chromosome 2, there are two peaks at 25 and 35 cM with the estimated effects $\hat{\alpha}_j^m = 0.2227$, $\hat{\alpha}_j^f = 0.2281$, and $\hat{\delta}_j = 0.1643$ and $\hat{\alpha}_j^m = 0.1832$, $\hat{\alpha}_j^f = 0.2517$, and $\hat{\delta}_j = 0.1829$. Obviously, it is difficult to distinguish one QTL or two QTL in chromosome 2 from the ML analysis.

The ML analyses of QTL mapping do not provide confidence intervals for the estimated QTL locations and effects. Confidence intervals would have to be determined by a resampling technique separately. We adopted the bootstrap method of Visscher *et al.* (1996b) to construct the confidence intervals for the estimated QTL locations. We used 1000 bootstrap samples to simulate the distributions of the locations (see Figure 1). The bootstrap means (the standard deviations) are 27.46 (10.08) cM and 25.19 (13.88) cM for design I, and 24.10 (13.57) cM and 32.2581 (27.71) cM for design II, for the two chromosomes, respectively.

In the MCMC analyses, we used the QTL intensity function of Sillanpää and Arjas (1998, 1999) to detect the number and locations of QTL. The interval length was chosen to be 1 cM long. The approximate posterior QTL intensities for both the binary and the normal data in design I are shown in Figure 2. The fact that the two approximate posterior QTL intensities both have three peaks around the true locations of the three simulated QTL supports a three-QTL model and the true model is indeed of three QTL in design I. Comparing the shapes of the QTL intensities for the binary and normal data analyses, we can see that binary data analysis does lose some information, but the information retained is still sufficient to detect all the simulated QTL.

Figure 3 depicts the approximate posterior QTL in-

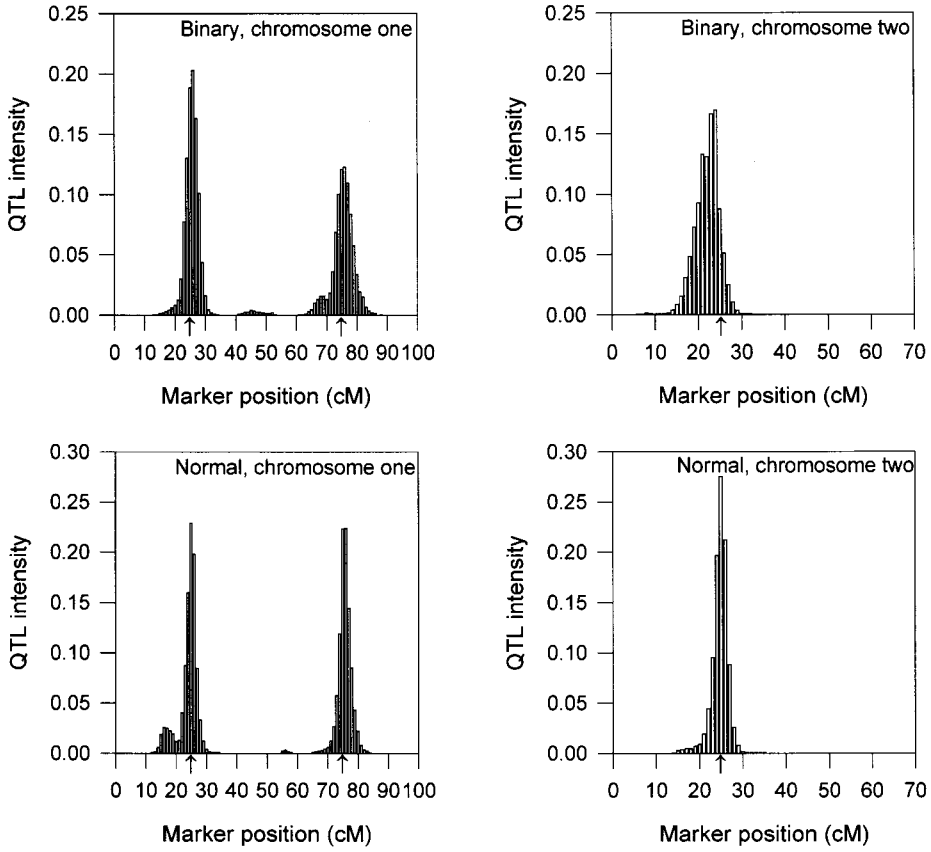

Figure 2.—Histograms of the posterior QTL intensity for binary data (top) and normally distributed data (bottom) in design I, respectively. Simulated true QTL locations are indicated with an arrow (↑).

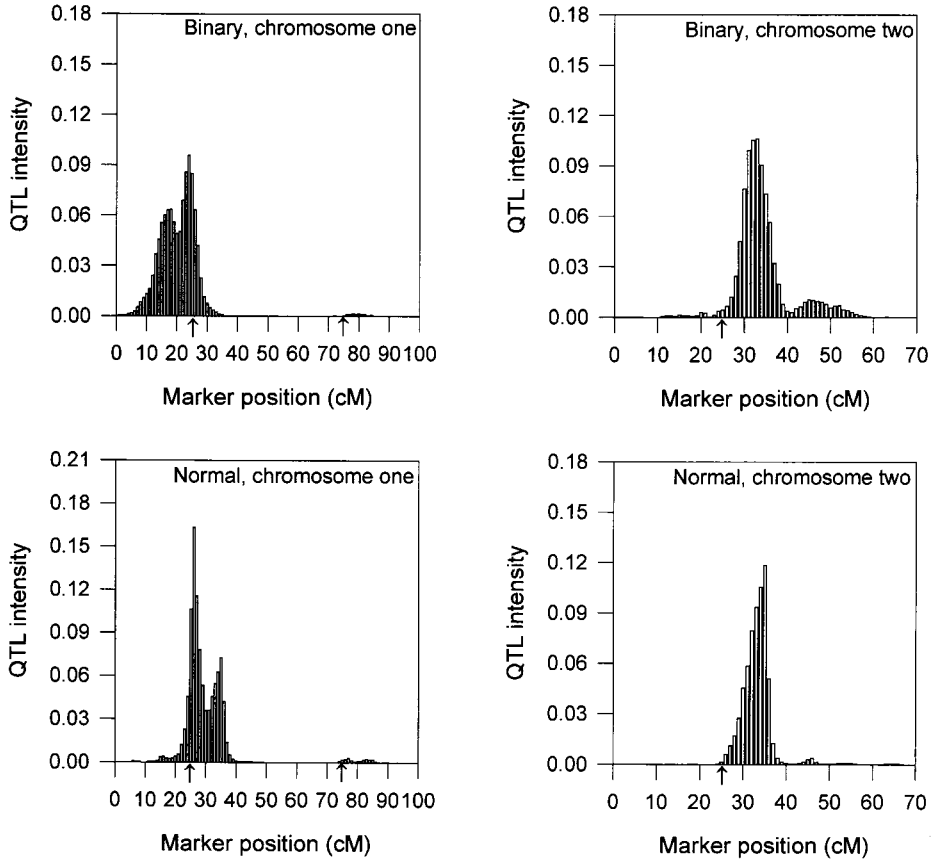

Figure 3.—Histograms of the posterior QTL intensity for binary data (top) and normally distributed data (bottom) in design II, respectively. Simulated true QTL locations are indicated with an arrow (↑).

TABLE 2

Empirical posterior distribution of QTL number and the posterior mean

Design	Data type	Estimated distribution, for l = 0	1	2	3	4	5	6	Estimated mean
I	Binary	0.0000	0.0142	0.1111	0.7622	0.1085	0.0040	0.0000	2.9771
	Normal	0.0000	0.0000	0.0073	0.9826	0.0101	0.0000	0.0000	3.0028
II	Binary	0.0000	0.1292	0.7713	0.0940	0.0056	0.0040	0.0000	1.9760
	Normal	0.0000	0.2319	0.7533	0.0142	0.0007	0.0000	0.0000	1.7839

tensities for both the binary and the normal data in design II. For chromosome 1, the QTL intensity graphs are concentrated around the first QTL locations for both the binary and the normal data. The second, the weakest QTL in chromosome 1, remained undetected for both the binary and the normal data, and this result was practically the same as that obtained from the ML analysis of the binary data in design II. For chromosome 2, the two approximate posterior QTL intensities both have one peak and apparently support one QTL residing at this chromosome. However, the modes of the intensities for the normal and binary data differ by ~9 cM from the simulated true location in chromosome 2.

The approximate posterior distributions for the number of QTL, obtained from the two designs, are presented in Table 2. For design I, the posterior means are essentially the same for the binary data and the normal data and coincide with the simulated number of QTL. As expected, the posterior variance of QTL number for the normal data is smaller than that for the binary data. Finally, the posterior mode of the QTL number overlaps with the true number for both types of data in design I. In the analysis of design II, the estimated posterior distributions for the number of QTL appear to have shifted to the left by 1 compared with the simulated number of QTL, for the two types of data. Again, as in design I, the posterior means and modes are essentially the same for both the binary data and the normal data.

Consider next the estimations of QTL effects. The estimates are reliable only in chromosome regions in which the posterior QTL intensity or the posterior density of QTL locations is sufficiently high (Sillanpää and Arjas 1998, 1999; Stephens and Fisch 1998). The highest posterior region attempts to capture a comparatively small region of the parameter space that contains most of the posterior probability mass. The chromosome regions with sufficiently high posterior QTL intensity are given in Table 3. We used only the posterior samples, in which QTL locations fall into the regions described in Table 3, to estimate the QTL effects. The posterior distributions of the QTL effects are presented graphically in Figure 4 for the three identified QTL for design I. The point estimates and the estimation errors of locations and effects of QTL for the two designs are presented in Table 3. In most cases the estimations of the QTL effects are reliable. As expected, the posterior variances of the normal data are smaller than those obtained for the binary data. For design I, the estimated QTL locations are very close to the corresponding true values and the standard errors are relatively small compared to design II, which has low heritabilities and skewed trait incidence.

DISCUSSION

We have presented here a Bayesian QTL mapping for complex binary traits. The methodology can be generalized to multiple-ordered categorical traits (appendix c). The most obvious advantage of the Bayesian method over existing ML is the ability to investigate the distributions of parameter estimates. Among the parameters of interest, the number of QTL may be the most important one. It is the Bayesian method that provides an easy way to estimate this parameter. The threshold model that the Bayesian mapping is based on is not new to QTL mapping for categorical traits. Bayesian mapping for normally distributed traits is also available. In this article we combine both techniques to develop the Bayesian mapping for binary traits. The main point of the threshold model is that by introducing an underlying normal variable into the problem, the binary response is connected to the normal linear model via the probit function. The major advantage of the threshold model as applied to Bayesian mapping is that once the underlying liability is generated, all other unknowns have conditional posterior distributions identical to those already given in Bayesian analysis of normal data.

Since the liability is a hypothetical variable, the interpretation of categorical data with a threshold model can be delicate. However, there are a number of ways to test the general validity of the model (Lynch and Walsh 1998, Chapter 25). In the threshold model, the liability is usually assumed normal. In reality, the nature of the underlying variable is always unknown. In Bayesian analysis, some kind of distribution must be assigned to this hypothetical variable and normal distribution is the natural choice. In addition, Tan *et al.* (1999) recently showed that the normal assumption for the liability is robust to obvious departure from normality.

The key focus of QTL mapping is on making infer-

TABLE 3

The highest posterior QTL intensity interval, Bayesian estimates of QTL locations, and allelic and dominance effects

Design	Data type	Chromosome	Interval (cM)	Sum of the QTL intensity	QTL location (cM)	α_j^m	α_j^f	δ_j
I	Binary	1	~20–30	0.7335	25.1873	0.4190	0.4498	0.5171
					(1.8353)	(0.1686)	(0.1800)	(0.1661)
		1	~67–82	0.7922	75.1281	0.1898	0.2563	0.3545
					(2.8636)	(0.1509)	(0.1432)	(0.1347)
		2	~19–30	0.8664	23.7155	0.1935	0.4041	0.4916
					(1.9447)	(0.2146)	(0.3352)	(0.3117)
	Normal	1	~21–28	0.8322	24.5080	0.3707	0.3131	0.4318
					(1.4152)	(0.0730)	(0.0737)	(0.1053)
		1	~70–80	0.9578	75.2191	0.1423	0.2708	0.3781
					(1.7838)	(0.0764)	(0.0688)	(0.0855)
		2	~21–28	0.9369	24.4426	0.3778	0.3079	0.4109
					(1.3540)	(0.0869)	(0.0766)	(0.0956)
II	Binary	1	~12–35	0.9311	22.1744	0.3422	0.2122	0.3334
					(4.7910)	(0.3251)	(0.2578)	(0.3946)
		2	~22–38	0.7383	32.8831	0.3205	0.2342	0.1226
					(2.9075)	(0.3593)	(0.2939)	(0.2131)
	Normal	1	~15–35	0.9822	28.0919	0.3362	0.2231	0.3184
					(4.0857)	(0.0798)	(0.0777)	(0.1475)
		2	~25–39	0.6330	33.3772	0.1321	0.2426	0.2972
					(2.4237)	(0.0659)	(0.0707)	(0.1328)

Posterior standard errors of the estimates are given in parentheses.

ences about the number of QTL, their locations, and effects. There are a number of advantages in arriving at inferential statements by using a Bayesian approach over the traditional methods. First, Bayes' method provides a complete posterior distribution for the number of QTL, their locations, and the corresponding effects. As a consequence, interval estimates of the parameters can be obtained straightforwardly. In contrast, ML only produces point estimates of these parameters. Confidence intervals would have to be determined separately, for example, by employing bootstrap (Visscher *et al.* 1996b) or other sampling-based methods. Second, it is usually difficult to determine the correct number of QTL using traditional methods. It has been shown that an incorrect specification for the number of QTL can lead to distortion of estimates of locations and effects when ML and least squares (LS) are used. Bayesian mapping allows one to include the number of QTL as an unknown in the analysis and thus avoids this distortion. Third, a common problem with traditional methods is how to choose the appropriate critical value of the statistical test for declaration of the presence of QTL. With the reversible jump MCMC, the number and the locations of QTL can be characterized by the posterior probability distribution of the number of QTL and the posterior QTL intensity. One can even calculate the posterior probability that some particular chromosomal region contains at least one QTL (Sillanpää and Arjas 1998, 1999). Finally, the Bayes method has the inherent flexibility introduced by its incorporation of multiple levels of randomness and the resultant ability to combine information from different sources. Therefore, the Bayesian approach could be extended to allow more complicated models for more complicated data structures.

An essential element of our full Bayesian mapping for binary traits is its ability to move between different values of the QTL number. The proposed method performed well for the simulated data. The mixing property of the MCMC algorithm does not seem to be overly sensitive to the choice of the initial values of the unknowns. For example, when we started with $l_0 = 6$, after 100 iterations l quickly dropped to 3 and subsequently behaved the same as when we started with $l_0 = 2$. A similar conclusion has been obtained by Heath (1997). However, the mixing property seems to be greatly affected by the proposal distribution for QTL effects when a new QTL is added to the model. Therefore, this proposal distribution should be chosen with care (Heath 1997; Stephens and Fisch 1998; Sillanpää and Arjas 1999). To ensure sufficient mixing, the single MCMC chain must be sufficiently long. Although the method is computationally very intensive, it does not need repeated analyses of resampled data, as required in ML for the permutation test. On a Sun SPARC 5 workstation, our analyses with a MCMC chain of 10^6 took ~6.5 hr for normal data and 8 hr for binary data, respectively. A major implementation issue in MCMC is to determine the effective sample sizes. This issue is related to the assessment of convergence of the MCMC sampler, the

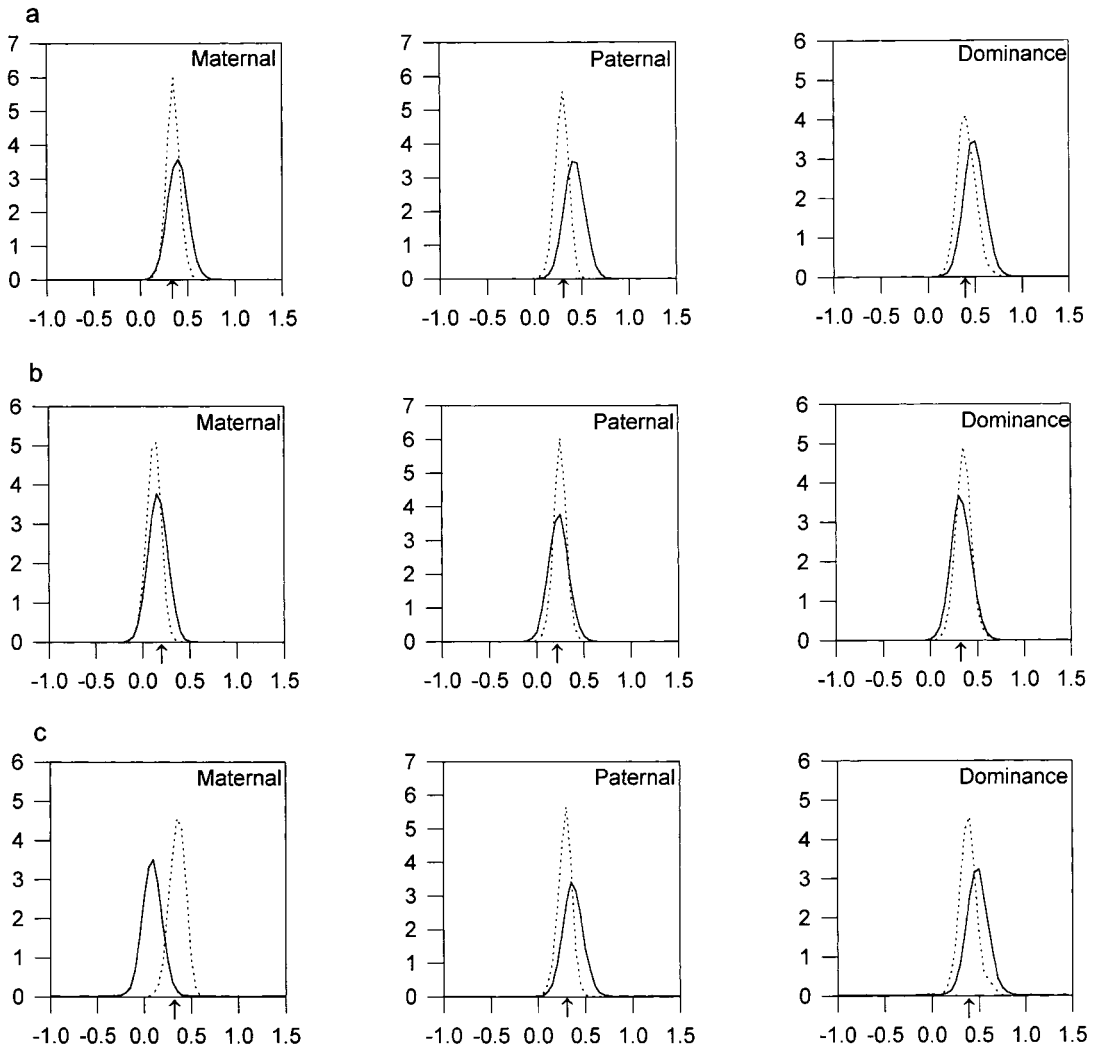

Figure 4.—Approximate posterior distributions of maternal allelic effects (α_j^f), paternal allelic effects (α_j^f), and dominance effects (δ_j) for $j = 1, 2, 3$, for design I. Simulated true values of QTL effects are indicated with an arrow (↑). The solid curves represent QTL mapping for binary data: (a) the first QTL determined from interval ~20–30 cM of chromosome 1; (b) the second QTL determined from interval ~67–82 cM of chromosome 1; (c) the QTL on chromosome 2 determined from interval ~19–30 cM. The dotted curves represent QTL mapping for normal data: (a) the first QTL determined from interval ~21–28 cM of chromosome 1; (b) the second QTL determined from interval ~70–80 cM of chromosome 1; (c) the QTL on chromosome 2 determined from interval ~21–28 cM.

serial correlation between the samples, and the burn-in period. When analyzing real data, one can examine time series graphs of simulated sequence and calculate the Monte Carlo variance to obtain estimates of the effective sample sizes of all parameters (Geyer 1992). In our simulation studies, it is difficult to calculate series correlation because the dimension keeps changing from one cycle to another. When the dimension changes, the identities of the QTL also change. Therefore, we empirically determined the burn-in period, the length of the MCMC chain, and the interval length of subsampling to reduce the serial correlation.

The Bayesian procedure presented in this study is based on known marker linkage phases in the parents. When the linkage phases are not known, they must be inferred first from marker genotypes of the parents and the offspring. If grandparents are also genotyped, the linkage phases can be accurately reconstructed; otherwise, a relatively large number of offspring for each family are required (Knott *et al.* 1996). Alternatively, one can treat linkage phases as random variables in the Bayesian analysis, as done by Sillanpää and Arjas (1999). When the family size is too small, inference of the parental linkage phases will be subject to large error and stochastic resampling is certainly required. When the mapping population contains many small families, accurate inferences of parental linkage phases are almost impossible and other statistical models may be considered, such as the IBD-based random model approach (Xu and Atchley 1995). Under the random model approach, one does not need to know the number of alleles and the parental linkage phases.

As an alternative to the approach of the liability augmentation, one can directly use the probit relationship between the binary phenotype and the model effects to simulate the model effects θ by using the Metropolis-Hastings algorithms, *i.e.*, replacing

$$p(\mathbf{S}|\mathbf{Y}, l, \lambda, \mathbf{Z}, \theta)p(\mathbf{Y}|l, \lambda, \mathbf{Z}, \theta)$$

by

$$p(\mathbf{S}|l, \lambda, \mathbf{Z}, \theta) = \int_{\mathbf{Y}} p(\mathbf{S}|\mathbf{Y}, l, \lambda, \mathbf{Z}, \theta)p(\mathbf{Y}|l, \lambda, \mathbf{Z}, \theta)d\,\mathbf{Y}.$$

In the simple situation, such as a single line cross or full-sib family, application of the probit model is possible. In fact, it will improve the mixing property of the MCMC because the unobserved liability has been integrated out. We decided to utilize the data augmentation approach by generating the liability because the method can be easily extended to more complicated situations, such as mapping for ordered categorical data or multiple traits.

As mentioned previously, the major advantage of Bayesian mapping is the ability to handle multiple QTL with multiple effects, including epistatic effects. Epistatic effects can be important in phenotypic evolution. Although we did not add the epistatic effects in our Bayesian model presented in this study, it is not difficult to do so. When a new QTL is proposed, its interactive effects with all existing QTL should be proposed and, if accepted, the epistatic effects should be included in the model. A QTL is finally added to the model if at least one effect caused by the QTL is accepted. This will involve additional reversible jumps on the dimension of the model even if the number of QTL remains the same.

Throughout the study, we update the genotype individual by individual and locus by locus. In general, this kind of single-site update does not always lead to an irreducible sampler because of strong dependency of close relatives and strong dependency of adjacent loci. In practice, when we deal with complicated pedigree data and closely linked markers, block updating more than one individual and more than one locus is required for appropriate mixing of the sampler (*e.g.*, Heath 1997; Sillanpää and Arjas 1999). In our study, because the pedigree is simple and the marker loci are not close enough, we did not find any insufficient mixing in genotype sampling. A Bayesian mapping for complicated pedigree data is now under investigation in this laboratory, where a block updating strategy is being incorporated in the sampler.

We thank Dr. D. D. Gessler for his helpful comments on the manuscript. We also thank two anonymous reviewers for their critical comments on an earlier version of the manuscript. This research was supported by the National Institutes of Health Grant GM55321-03 and the U.S. Department of Agriculture National Research Initiative Competitive Grants Program 97-35205-5075 to S.X.

LITERATURE CITED

Albert, J. H., and S. Chib, 1993 Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88:** 669–679.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Devroye, L., 1986 *Non-Uniform Random Variable Generation*. Springer-Verlag, New York.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longman, London.

Gelfand, A. E., and A. F. M. Smith, 1990 Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85:** 398–409.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis.* Chapman & Hall, London.

Geyer, C. J., 1992 Practical Markov chain Monte Carlo. Stat. Sci. **7:** 473–511.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

Hackett, C. A., and J. I. Weller, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics **51:** 1252–1263.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.

Hoeschele, I., and P. VanRanden, 1993a Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theor. Appl. Genet. **85:** 953–960.

Hoeschele, I., and P. VanRanden, 1993b Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor. Appl. Genet. **85:** 946–952.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

Jansen, R. C., D. L. Johnson and J. A. M. Van Arendonk, 1998 A mixture approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. Genetics **148:** 391–399.

Kao, C. H., Z.-B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

Knott, S. A., J. M. Elsen and C. S. Haley, 1996 Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. Theor. Appl. Genet. **93:** 71–80.

Kruglyak, L., and E. S. Lander, 1995 A nonparametric approach for mapping quantitative trait loci. Genetics **139:** 1421–1428.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

Rao, S., and S. Xu, 1998 Mapping quantitative trait loci for ordered categorical traits in four-way crosses. Heredity **81:** 214–224.

Rebai, A., 1997 Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. Genet. Res. **69:** 69–74.

Richardson, S., and P. J. Green, 1997 On Bayesian analysis of mixtures with an unknown number of components. J. R. Stat. Soc. Ser. B **59:** 731–792.

Satagopan, J. M., and B. S. Yandell, 1996 Estimating the number of quantitative trait loci via Bayesian model determination. Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Diseases, Biometric Section, Joint Statistical Meeting, Chicago.

Satagopan, J. M., B. S. Yandell, M. A. Newton and T. C. Osborn, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics **144:** 805–816.

Sillanpää, M. J., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics **148:** 1373–1388.

Sillanpää, M. J., and E. Arjas, 1999 Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics **151:** 1605–1619.

Sorensen, D. A., S. Andersen, D. Gianola and I. Korsgaard, 1995 Bayesian inference in threshold models using Gibbs sampling. Genet. Sel. Evol. **27:** 229–249.
Stephens, D. A., and R. D. Fisch, 1998 Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics **54:** 1334–1347.
Tan, M., Y. Qu and J. S. Rao, 1999 Robustness of the latent variable model for correlated binary data. Biometrics **55:** 258–263.
Tanner, M. A., and W. H. Wong, 1987 The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82:** 528–549.
Thaller, G., and I. Hoeschele, 1996 A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: I. Methodology. Theor. Appl. Genet. **93:** 1161–1166.
Uimari, P., and I. Hoeschele, 1997 Mapping linked quantitative trait loci using Bayesian method analysis and Markov chain Monte Carlo Algorithms. Genetics **146:** 735–743.
Uimari, P., G. Thaller and I. Hoeschele, 1996 The use of multiple markers in a Bayesian method for mapping quantitative trait loci. Genetics **143:** 1831–1842.
Visscher, P. M., C. S. Haley and S. A. Knott, 1996a Mapping QTL for binary traits in backcross and F_2 populations. Genet. Res. **68:** 55–63.
Visscher, P. M., R. Thomson and C. S. Haley, 1996b Confidence intervals in QTL mapping by bootstrapping. Genetics **143:** 1013–1020.
Xu, S., and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics **141:** 1189–1197.
Xu, S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics **143:** 1417–1424.
Xu, S., N. Yonash, R. L. Vallejo and H. H. Cheng, 1998 Mapping quantitative trait loci for complex binary traits using a heterogeneous residual variance model: an application to Marek's disease susceptibility in chickens. Genetica **104:** 171–178.
Yi, N., and S. Xu, 1999a Mapping quantitative trait loci for complex binary traits in outbred populations. Heredity **82:** 668–676.
Yi, N., and S. Xu, 1999b A random approach to mapping quantitative trait loci for complex binary traits in outbred populations. Genetics **153:** 1029–1040.
Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: T. F. C. Mackay

APPENDIX A: CONDITIONAL POSTERIOR DISTRIBUTIONS

Conditional posterior distribution of the liability y_i: Conditional on θ, $\mathbf{Z}_i$, and s_i, the liability y_i follows a truncated normal distribution. Depending on the binary phenotypic value s_i, we have

$$p(y_i|\theta, \mathbf{Z}_i, s_i = 1) = \frac{\varphi(y_i - \mathbf{X}_i^T\beta - \sum_{j=1}^{l} \mathbf{Z}_{ij}^T\mathbf{H}\gamma_j, 1)}{\Phi(\mathbf{X}_i^T\beta + \sum_{j=1}^{l}\mathbf{Z}_{ij}^T\mathbf{H}\gamma_j)}1(y_i > 0) \tag{A1}$$

and

$$p(y_i|\theta, \mathbf{Z}_i, s_i = 0) = \frac{\varphi(y_i - \mathbf{X}_i^T\beta - \sum_{j=1}^{l} \mathbf{Z}_{ij}^T\mathbf{H}\gamma_j, 1)}{1 - \Phi(\mathbf{X}_i^T\beta + \sum_{j=1}^{l}\mathbf{Z}_{ij}^T\mathbf{H}\gamma_j)}1(y_i \le 0), \tag{A2}$$

where $\varphi(x, \sigma^2)$ is the normal density with mean zero and variance σ^2 and $\Phi(\cdot)$ is the standardized normal distribution function.

Conditional posterior distributions of the regression coefficients: Given the liability $\mathbf{Y}$, the number l, and the genotype $\mathbf{Z}$ of QTL, the posterior distribution for θ can be derived using the standard normal linear model theory. If normal priors are chosen for the regression coefficients θ, these posterior distributions are given by

$$\beta_k|\mathbf{Y}, l, \mathbf{Z}, \{\beta_j\}_{j\neq k}, \gamma \sim N\left(\frac{\beta_{0k}/\sigma^2_{\beta_k} + \sum_{i=1}^{n} x_{ik}(y_i - \mathbf{X}_i^T\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^T\mathbf{H}\gamma_i + x_{ik}\beta_k)}{1/\sigma^2_{\beta_k} + \sum_{i=1}^{n} x_{ik}^2}, \frac{1}{1/\sigma^2_{\beta_k} + \sum_{i=1}^{n} x_{ik}^2}\right), \tag{A3}$$

for $k = 1, \cdots, p$. Each of the allelic effects has a normal posterior distribution,

$$\alpha_{j*}^{v}|\mathbf{Y}, l, \mathbf{Z}, \beta, \{\gamma_j\}_{j\neq j*}, \alpha_{j*}^{\bar{v}}, \delta_{j*} \sim N\left(\left[\frac{\alpha_{0j*}^{v}}{\sigma^2_{\alpha_{j*}^{v}}} + \sum_{i=1}^{n}(\mathbf{Z}_{ij*}^T\mathbf{H}_v)[y_i - X_i^T\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^T\mathbf{H}\gamma_j + (\mathbf{Z}_{ij*}^T\mathbf{H}_v)\alpha_{j*}^{v}]\right]/\left[\frac{1}{\sigma^2_{\alpha_{j*}^{v}}} + \sum_{i=1}^{n}(\mathbf{Z}_{ij*}^T\mathbf{H}_v)^2\right], 1/\left[\frac{1}{\sigma^2_{\alpha_{j*}^{v}}} + \sum_{i=1}^{n}(\mathbf{Z}_{ij*}^T\mathbf{H}_v)^2\right]\right), \tag{A4}$$

where $v = m$, s and

$$\bar{v} = \begin{cases} m & \text{if } v = f \\ f & \text{if } v = m \end{cases}.$$

The posterior distribution of the dominance effect is

$$\delta_{j*}|\mathbf{Y}, l, \mathbf{Z}, \beta, \{\gamma_j\}_{j\neq j*}, \alpha_{j*}^{m}, \alpha_{j*}^{f} \sim N\left(\left[\frac{\delta_{0j*}}{\sigma^2_{\delta_{j*}}} + \sum_{i=1}^{n}(\mathbf{Z}_{ij*}^T\mathbf{H}_\delta)[y_i - \mathbf{X}_i^T\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^T\mathbf{H}\gamma_j + (\mathbf{Z}_{ij*}^T\mathbf{H}_\delta)\delta_{j*}]\right]/\left[\frac{1}{\sigma^2_{\delta_{j*}}} + \sum_{i=1}^{n}(\mathbf{Z}_{ij*}^T\mathbf{H}_\delta)^2\right], 1/\left[\frac{1}{\sigma^2_{\delta_{j*}}} + \sum_{i=1}^{n}(\mathbf{Z}_{ij*}^T\mathbf{H}_\delta)^2\right]\right), \tag{A5}$$

for $j* = 1, \cdots, l$, where β_{0k}, α_{0j*}^{m}, α_{0j*}^{f}, and δ_{0j*} are the prior means for β_k, α_{j*}^{m}, α_{j*}^{f}, and δ_{j*}, respectively, and $\sigma^2_{\beta_k}$, $\sigma^2_{\alpha_{j*}^{m}}$, $\sigma^2_{\alpha_{j*}^{f}}$, and $\alpha^2_{\delta_{j*}}$ are the prior variances for β_k, α_{j*}^{m}, α_{j*}^{f}, and δ_{j*}.

Conditional posterior distributions of the QTL and marker genotypes: The conditional posterior distribution of the QTL genotype $\mathbf{Z}_{ij}$ is a discrete distribution over the possible genotypes. In model (1), the QTL genotype $\mathbf{Z}_{ij}$ takes one of four values. Thus, for instance, the conditional posterior distribution that an individual takes a genotype $z_{ij} = (1, 0, 0, 0)^T$ is given by

$$p(\mathbf{Z}_{ij} = z_{ij}|y_i, \lambda, \mathbf{Z}_{i(-j)}, \mathbf{M}_i, \theta) = \frac{p(y_i|\lambda, \mathbf{Z}_{ij} = z_{ij}, \mathbf{Z}_{i(-j)}, \theta)p(\mathbf{Z}_{ij} = z_{ij}|\lambda_j, \mathbf{Z}_{ij}^l, \mathbf{Z}_{ij}^r)}{\sum_{\mathbf{Z}_{ij}} p(y_i|\lambda, \mathbf{Z}_{ij}, \mathbf{Z}_{i(-j)}, \theta)p(\mathbf{Z}_{ij}|\lambda_j, \mathbf{Z}_{ij}^l, \mathbf{Z}_{ij}^r)}, \tag{A6}$$

where $\mathbf{Z}_{i(-j)} = \{\mathbf{Z}_{ij'} : 1 \le j' \le l, j' \neq j\}$, and $\mathbf{Z}_{ij}^l$ ($\mathbf{Z}_{ij}^r$) is the left (right) flanking genotype of the jth QTL of the ith individual (markers or QTL).

The conditional posterior distribution of the marker

genotype $\mathbf{M}_{ij}$ is dependent only on the genotypes of relevant flanking loci (marker or QTL). Taking the prior of the marker into consideration, we can obtain

$$p(\mathbf{M}_{ij} = m_{ij}|\mathbf{M}^{l}_{ij},\ \mathbf{M}^{r}_{ij}) = \frac{p(\mathbf{M}_{ij} = m_{ij})p(\mathbf{M}^{l}_{ij}, \mathbf{M}^{r}_{ij}|\mathbf{M}_{ij} = m_{ij})}{p(\mathbf{M}^{l}_{ij}, \mathbf{M}^{r}_{ij})}, \tag{A7}$$

where $\mathbf{M}^{l}_{ij}$ ($\mathbf{M}^{r}_{ij}$) is the left (right) flanking complete genotype of the jth marker of the ith individual, and $p(\mathbf{M}_{ij} = m_{ij})$ is the prior probability of the jth marker of the ith individual. $p(\mathbf{M}_{ij} = m_{ij})$ is calculated by the multipoint method (Rao and Xu 1998).

APPENDIX B: ACCEPTANCE PROBABILITIES

Updating QTL locations: To modify the location λ_j of the jth QTL, a proposal λ_j^{new} is generated from a uniform distribution on the interval $[\lambda_j - d, \lambda_j + d]$, where d is a tuning parameter. The acceptance probability for the change from λ_j to λ_j^{new} takes $\min\{1, \alpha\}$, where the relative importance ratio α is defined as

$$\alpha = \frac{p(\mathbf{Y}, l, \lambda_j^{new}, \lambda_{-j}, \mathbf{M}, \mathbf{Z}, \theta|\mathbf{S})}{p(\mathbf{Y}, l, \lambda_j, \lambda_{-j}, \mathbf{M}, \mathbf{Z}, \theta|\mathbf{S})} = \prod_{i=1}^{n}\frac{p(\mathbf{Z}_{ij}|\lambda_j^{new}, \mathbf{Z}^{l*}_{ij}, \mathbf{Z}^{r*}_{ij})}{p(\mathbf{Z}_{ij}|\lambda_j, \mathbf{Z}^{l}_{ij}, \mathbf{Z}^{r}_{ij})}, \tag{B1}$$

where $\lambda_{-j} = \{\lambda_{j'} : 1 \le j' \le l, j' \ne j\}$, $\mathbf{Z}^{l}_{ij}$ ($\mathbf{Z}^{r}_{ij}$) is the genotype of the left (right) flanking locus (marker or QTL) at the current position for the jth QTL in the ith individual, and $\mathbf{Z}^{l*}_{ij}$ ($\mathbf{Z}^{r*}_{ij}$) is the genotype of the left (right) flanking locus at the proposed new location for the jth QTL in the ith individual.

Add one new QTL: Given the liability $\mathbf{Y}$, the acceptance probability is the same as that of the normal trait, except that the normal observables are replaced by the simulated values of liability. As in Sillanpää and Arjas (1998), the acceptance probability is $\min\{1, \alpha\}$, where

$$\alpha = \frac{\exp\{-\tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{X}_i^{T}\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^{T}\mathbf{H}\gamma_j - \mathbf{Z}_i^{*T}\mathbf{H}\gamma^{*})^2\}}{\exp\{-\tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{X}_i^{T}\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^{T}\mathbf{H}\gamma_j)^2\}} \times \frac{\mu}{l+1} \times \frac{p_d}{(l+1)p_a}, \tag{B2}$$

where $\mathbf{Z}_i^{*}$ is the proposed genotype of the ith individual, γ^{*} are the proposed QTL effects, and μ is the prior mean of the QTL number.

Delete one QTL: If the jth existing QTL is proposed to be removed from the model, the acceptance probability is $\min\{1, \alpha\}$, where

$$\alpha = \frac{\exp\{-\tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{X}_i^{T}\beta - \sum_{j'\ne j}^{l}\mathbf{Z}_{ij'}^{T}\mathbf{H}\gamma_{j'})^2\}}{\exp\{-\tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{X}_i^{T}\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^{T}\mathbf{H}\gamma_j)^2\}} \times \frac{l}{\mu} \times \frac{lp_a}{p_d}. \tag{B3}$$

In (B2) and (B3), the first term is the likelihood ratio, the second term is the prior ratio, and the third term is the proposal ratio; Jacobian is 1.

APPENDIX C: GENERALIZATION TO MULTIPLE-ORDERED CATEGORICAL TRAITS

The method described in the text can be generalized to multiple-ordered categorical traits. Suppose now that the observed phenotypic value s_i takes one of c ordered categories, $1, \cdots, c$. A set of fixed thresholds, $t_1 < t_2 < \cdots < t_{c-1}$, in the scale of the liability determine the observed categories. Let $t_0 = -\infty$ and $t_c = +\infty$. We observe s_i, where $s_i = k$ if $t_{k-1} < y_i \le t_k$ ($k = 1, \cdots, c$). The thresholds $t_1, \cdots, t_{c-1}$ are unknown and need to be estimated. To ensure that the parameters are identifiable, it is necessary to impose one restriction on the thresholds. Without loss of generality, we take $t_1 = 0$ (Albert and Chib 1993) and estimate $\mathbf{t} = (t_2, \cdots, t_{c-1})$. Assuming that t and (l, λ, θ) are independently distributed *a priori*, the joint posterior distribution can be written as

$$p(\mathbf{Y}, l, \lambda, \mathbf{M}, \mathbf{Z}, \theta, \mathbf{t}\ |\mathbf{S}) \propto p(\mathbf{S}|\mathbf{Y}, l, \lambda, \mathbf{Z}, \theta, \mathbf{t})p(\mathbf{Y}|l, \lambda, \mathbf{Z}, \theta) \times p(\mathbf{Z}|l, \lambda, \mathbf{M})p(\mathbf{M})p(l, \lambda, \theta)p(\mathbf{t}), \tag{C1}$$

where $p(\mathbf{Y}|l, \lambda, \mathbf{Z}, \theta)$, $p(\mathbf{Z}|l, \lambda, \mathbf{M})$, $p(\mathbf{M})$, and $p(l, \lambda, \theta)$ are the same as for the binary data model. The first term in (C1) is the likelihood and expressed as (Albert and Chib 1993; Sorensen *et al.* 1995)

$$p(\mathbf{Y}|l, \lambda, \mathbf{Z}, \theta) = \prod_{i=1}^{n}\left\{\sum_{k=1}^{c}1(t_{k-1} < y_i \le t_k)1(s_i = k)\right\}. \tag{C2}$$

The last term in (C1), $p(\mathbf{t})$, is the prior density of $\mathbf{t}$ and is discussed below.

It is clear that the conditional posterior distributions of θ, $\mathbf{M}$, and $\mathbf{Z}$ are the same as those specified in the binary model. For the liability associated with the kth observation, we have

$$p(y_i|\theta, \mathbf{Z}_i, s_i = k) = \frac{\varphi(y_i - \mathbf{X}_i^{T}\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^{T}\mathbf{H}\gamma_j, 1)}{\Phi(t_k - \mathbf{X}_i^{T}\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^{T}\mathbf{H}\gamma_j) - \Phi(t_{k-1} - \mathbf{X}_i^{T}\beta - \sum_{j=1}^{l}\mathbf{Z}_{ij}^{T}\mathbf{H}\gamma_j)}1 \quad (t_{k-1} < y_i \le t_k), \tag{C3}$$

where $\varphi(x, \sigma^2)$ stands for the normal density with mean 0 and variance σ^2, and $\Phi(\cdot)$ is the standardized normal distribution function. If we assign a diffuse prior for $\mathbf{t}$, the conditional posterior distribution of t_k given $\{t_j, j \ne k\}$ and all the other parameters is uniform on the interval $[\max\{\max\{y_i : s_i = k\}, t_{k-1}\}, \min\{\min\{y_i : s_i = k+1\}, t_{k+1}\}]$ (Albert and Chib 1993; Sorensen *et al.* 1995).

The MCMC algorithm for the binary case described in the text is now generalized to multiple-ordered categorical traits. In brief, we only need to modify the likelihood and generated $\mathbf{t}$. Eventually, the liability is generated according to a doubly truncated normal rather than a singly truncated one. Updating of other parameters remains the same as in binary data analysis.